Sentiment Analysis of Post-Approval Birth Control Reviews

Hayley Zorkic
HLZ227
Hayley.zorkic@utexas.edu

Karinne Berstis
KB227
email@domain

Abstract

In this project, we are interested in exploring a variety of models that can classify positive, negative, and neutral public sentiment of post-approval drug reviews, specifically Birth Control. There are formal channels for which biotechnology companies, biopharmaceutical companies, and regulatory agencies (BBRAs) can gain an understanding for the safety and efficacy of drugs such as medical records data or post-approval investigations, however we believe there is a wealth of knowledge to be found in the informal reviews of drugs. We will explore

1 Introduction

Stevens (2018) showed that there is evidence that medical providers may diminish patient's concerns regarding adverse side effects to contraceptives. In these cases, patients may turn to social media resources to verify and report their experiences–positive and negative. Additionally, Wartella, et al. (2016) showed that social media is increasingly used by young adults to report health issues and seek health information. Tweets, reddit posts, or other drug review websites may provide valuable insight into the nature of an individual's experience with a particular pharmaceutical, as well as indicate if negative side effects are more widespread than originally indicated by trials. Generalized, these models will allow biotechnology, biopharmaceutical, and regulatory agencies to gain a better understanding on how drugs are being received in a real-world context which may remove some bias caused by the formal trial setting.

2 Data and Resources

A dataset of drug reviews from the UCI Machine Learning Repository is linked here. The dataset provides short response patient reviews on specific drugs, the related conditions and a 10 star patient rating reflecting overall patient satisfaction with the drug. The data was obtained by crawling online pharmaceutical review sites (Gräßer, 2018). The dataset includes 215,063 reviews in total, 28,930 of which are for Birth Control. We could split the data in the following ways:

Train	Development	Test
All UCI Drugs	All UCI Drugs	BC in UCI
All UCI Drugs	All UCI Drugs	Scraped BC
BC in UCI	BC in UCI	Scraped BC
BC in UCI	BC in UCI	BC in UCI

Where the training set is what we use to train out models, the dev set is what we use to test and select which models to proceed with, and our test set is what we use to see how good our model is on new/unseen data. If we choose to scrape new birth control reviews for our test set, we can scrape from pharmaceutical review sites, reddit posts or tweets (using the appropriate python packages). If we select tweets or reddit posts, we will hand label a handful of reviews so we can evaluate the performance of our models.

3 Methodology

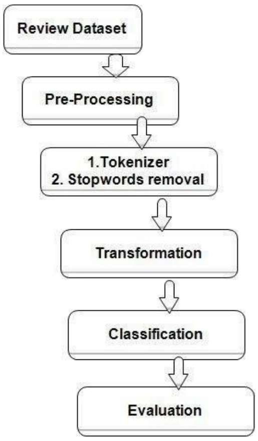
There are several models that can be used for sentiment analysis. We would like to apply some models discussed in class as well as some models said to be good for text analysis which were outlined in the literature (Vijayaraghavan, 2020; aiperspectives.com, 2021):

1. Naive Bayes
2. Logistic Regression
3. Random Forest
4. k-Nearest Neighbors
5. Artificial Neural Networks (ANN) (perhaps Recurrent Neural Networks with Long Short Term Memory)

4 Evaluation

Sentiment Analysis uses the evaluation metrics of Precision, Recall, F-score, and Accuracy, depending on the balance of classes in our dataset, we can choose the appropriate metric. Some common metrics that we can use are:

1. ROC AUC and Accuracy Score if our data is balanced
2. Precision-Recall AUC and F1 scores if our data is unbalanced



5 Team Structure

Karinne:

1. Clean and pre-process UCI data. This includes downloading, transforming variables, etc.
2. Visualize model performance with AUC/ROC or Precision/Recall curve? Word cloud for positive and negative reviews?
3. Results: Describe as clearly as possible what your system can (and cannot) do. You can show examples of things your system is getting correct and of errors it is making. If applicable, measure performance by some performance measure.

Hayley:

1. Obtain birth control reviews from the internet- scrape them off some website maybe? Alternatively, we could choose a type of drug already in the UCI dataset
2. Documentation of the code, but we can do this as we go along
3. Check code with Jessy
4. Evaluate Classifiers. Decide which metric to use and when. Do they have different metrics?

Both:

1. Coding everything. We have a skeleton from a lot of our assignments; however, we will need to adapt it. Pair-programming.
2. Create the final project report 3-4 pages.

6 Citations

12 twitter sentiment analysis algorithms compared. AI Perspectives. (2021, June 15). Retrieved March 25, 2022, from https://www.aiperspectives.com/twitter-sentiment-analysis/

5 things you need to know about sentiment analysis and classification. KDnuggets. (n.d.). Retrieved March 25, 2022, from https://www.kdnuggets.com/

Pushshift. (n.d.). *Pushshift/API: Pushshift API*. GitHub. Retrieved March 25, 2022, from https://github.com/pushshift/api

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018, April). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health* (pp. 121-125).

Romer, D. Adolescents in the digital age: Effects on health and development. Media Commun. 2016; 4 (3): 1-94. ISSN: 2183-2439. doi: 10.17645/mac. v4i3. 659.

Stevens, L. M. (2018). "We have to be mythbusters": Clinician attitudes about the legitimacy of patient concerns and dissatisfaction with contraception. *Social Science & Medicine*, *212*, 145-152.

Tweepy. (n.d.). Retrieved March 25, 2022, from https://www.tweepy.org/

Twitter. (n.d.). *Twitter API for academic research | products | twitter developer platform*. Twitter. Retrieved March 25, 2022, from https://developer.twitter.com/

Vijayaraghavan, S., & Basu, D. (2020). Sentiment analysis in drug reviews using supervised machine learning algorithms. *arXiv preprint arXiv:2003.11643*. https://arxiv.org/pdf/2003.11643v1.pdf

Wickham, H. (2020, July 20). *Tools for working with urls and HTTP [R package httr version 1.4.2]*. The Comprehensive R Archive Network. Retrieved March 25, 2022, from https://cran.r-project.org/web/packages/httr/index.html?Author=Andrew%2520Pitre&Preview=true